Adaptive PFC Headroom

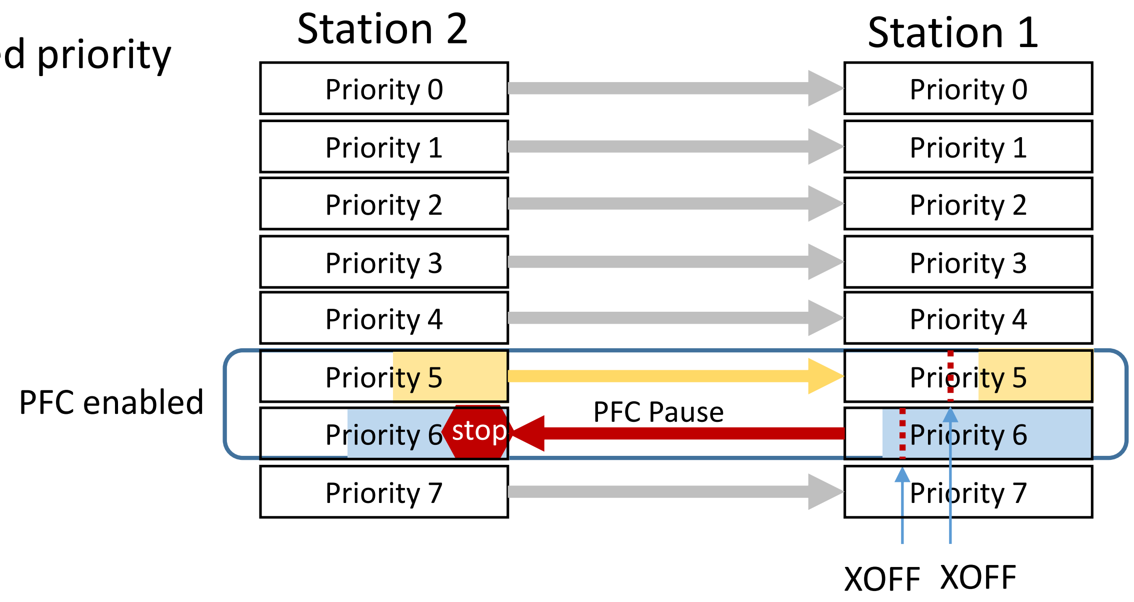
Lily Lv (Huawei)

Outline

- PFC Recap : Accurate ‘Headroom’ is Important for PFC
- Complexity of Headroom Calculation
- What We Have in IEEE 802
- What is Still Missing
- Proposal for Adaptive PFC Headroom
- Next Steps

Recap: PFC Concept

- Priority based Flow Control (PFC) is defined in Clause 36 of IEEE Std 802.1Q-2018
  - Mainly used in data center networks in order to avoid packet loss due to congestion.
  - “PFC allows link flow control to be performed on a per-priority basis. In particular, PFC is used to inhibit transmission of data frames associated with one or more priorities for a specified period of time. PFC can be enabled for some priorities on the link and disabled for others.” (Std 802.1Q-2018)
- One example of PFC
  - XOFF threshold which invokes PFC is set on each PFC enabled priority
  - Priority 6 reaches threshold XOFF
  - PFC pause frame is triggered and sent upstream
  - Upstream priority 6 transmission is stopped



Recap: PFC Delay and Headroom

- There is a time delay between PFC invocation on sender and pause action on receiver.
- This PFC delay requires the PFC sender (station 1 in figure) reserve buffer to absorb in-flight packets.
- The reserved buffer is also called 'headroom'.

Figure N-1 provides an high-level view of the various delays to consider:

a) Processing and queuing delay of the PFC request
b) Propagation delay of the PFC frame across the media
c) Response time to the PFC indication at the far end
d) Propagation delay across the media on the return path

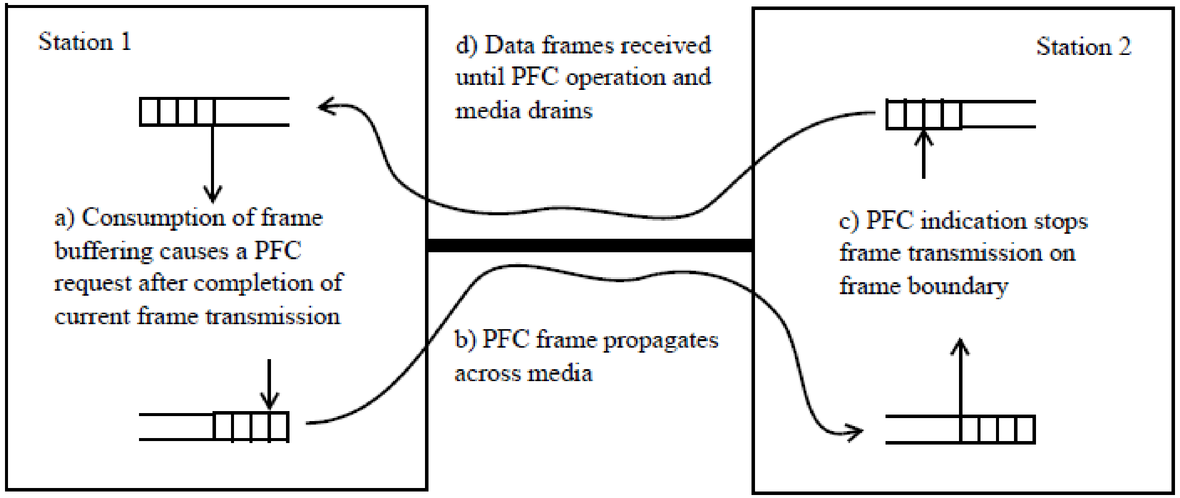
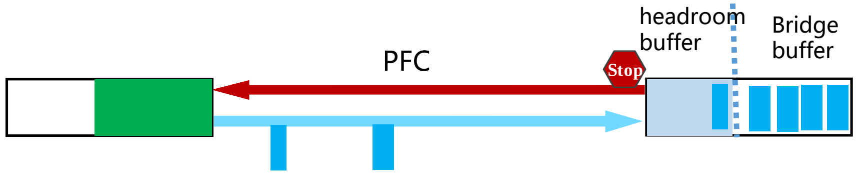


Figure N-1—PFC delays

Accurate 'Headroom' is Important for PFC

- PFC headroom and bridge buffers share the same buffer pool in many implementations.
  - When PFC is not invoked, bridge buffer is used to absorb normal traffic burst
  - When PFC is invoked, buffers as PFC Headroom is used to absorb inflight packets

Bridge buffer is used to absorb burst traffic, to avoid packet loss.



PFC must reserve headroom buffer to absorb inflight packet, to avoid packet loss.

- XOFF threshold setting relates to headroom.
  - If XOFF threshold is too high (less headroom)
    - packet drop may happen, not 'lossless' anymore.
  - If XOFF threshold is too low (more headroom)
    - traffic is suspended unnecessarily, low network bandwidth utilization.
    - Buffer resource is wasted.
- By calculating headroom, optimal XOFF threshold could be set.

Complexity of Headroom Calculation

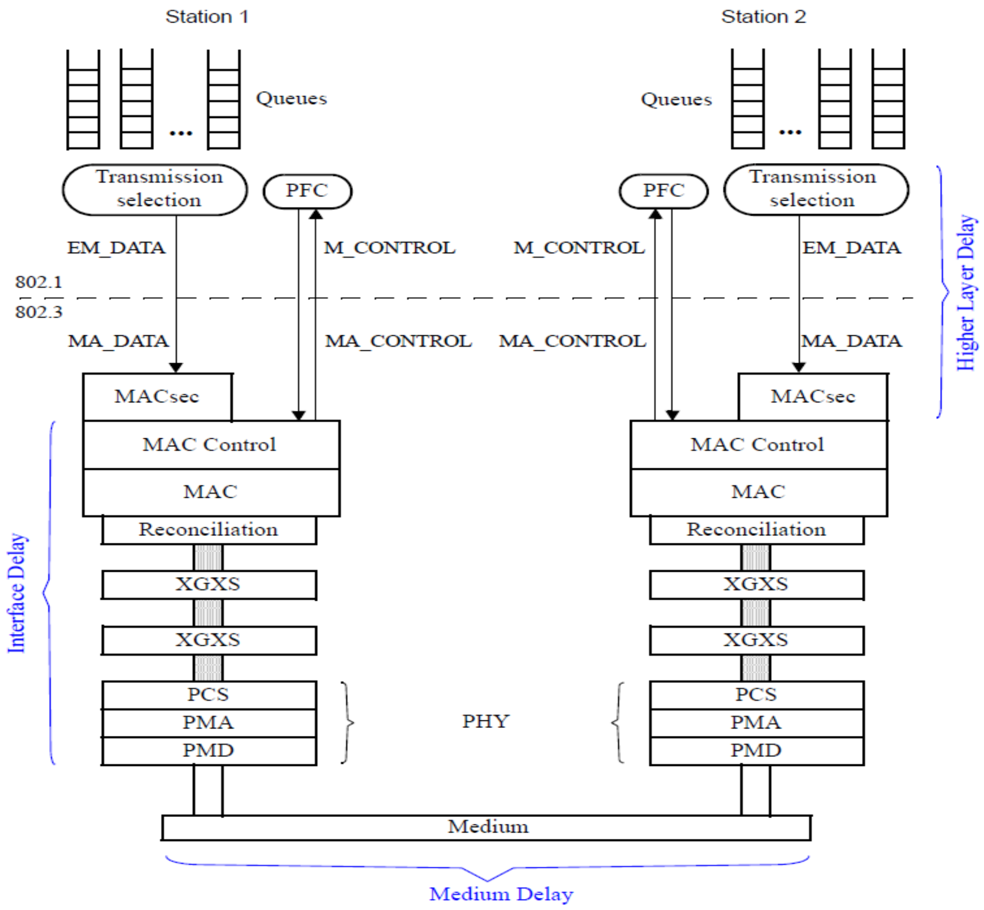


Figure N-2—Delay model (802.1Q-2018)

- Delay value calculation:
  - PFC transmission delay need consider maximum length frame as worst case, as well as PFC frame itself.
  - Interface delay and higher layer delay are vendor implementation dependent
    - 802.3 defines maximum value of such delays, however, vendors can do much better than that.
  - Medium delay is port speed, media and distance dependent

Complexity of Headroom Calculation

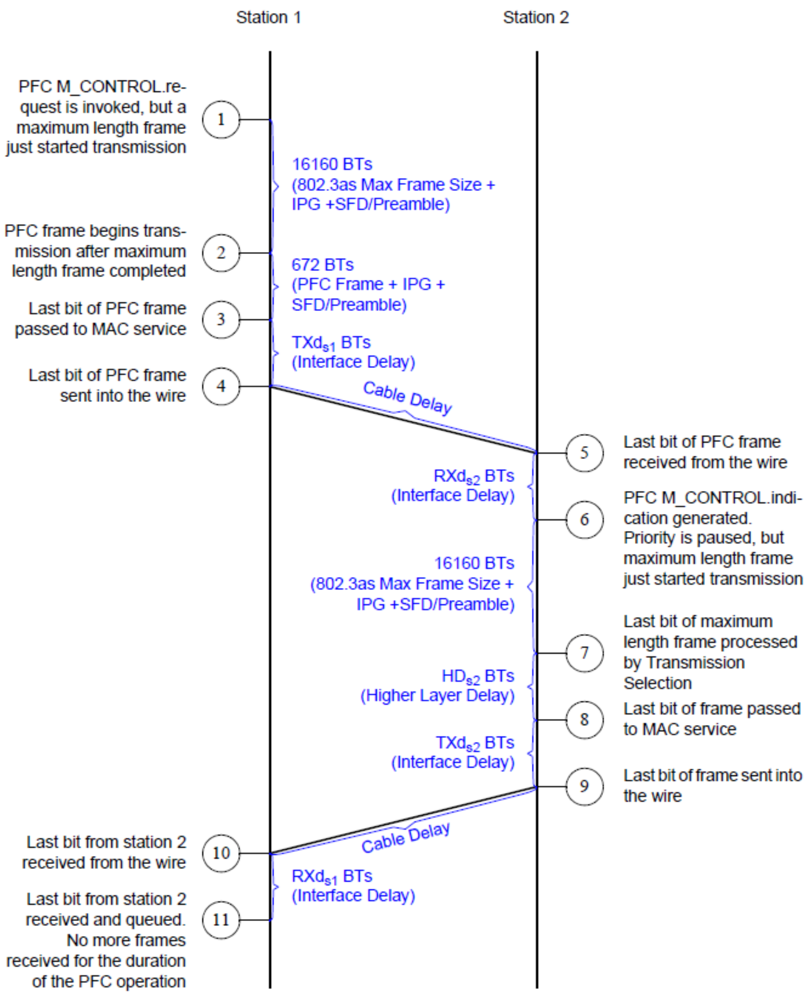


Figure N-3—Worst-case delay (802.1Q-2018)

- One example from 802.1Q-2018 Annex N, assuming 100 meters cat6 cable, 10G BASE-T with maximum interface delay and higher layer delay.

  DV = 2 ╳ (Max Frame) + (PFC Frame) + 2 ╳ (Cable Delay) + (TXds1 + RXds2) + (TXds2 + RXds1) + HDs2

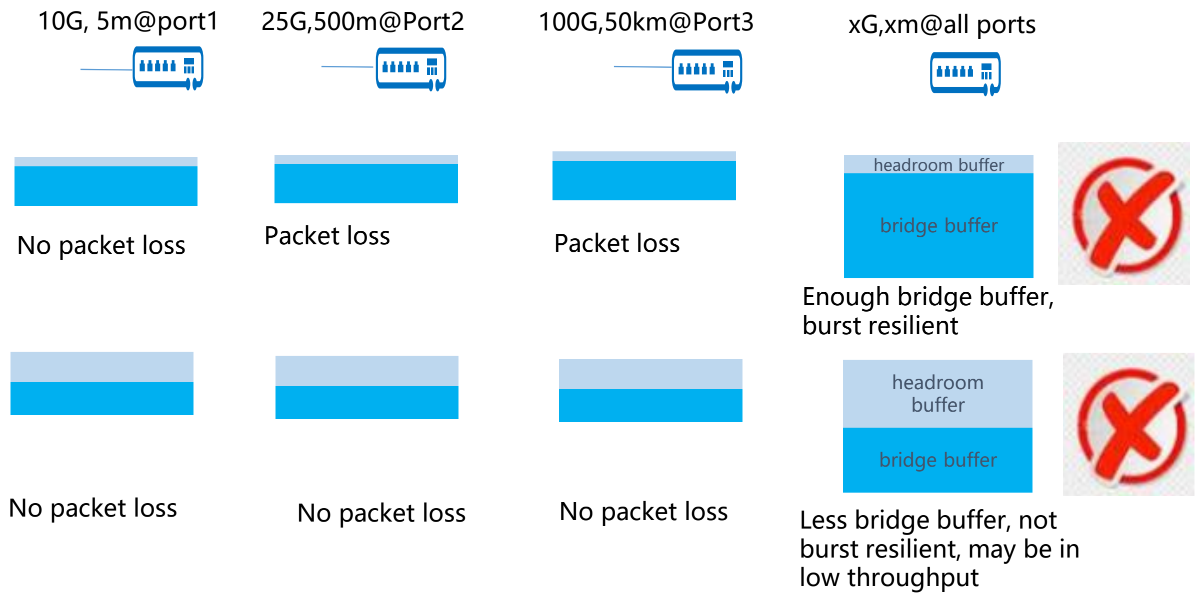
  DV = 2 ╳ (16 160) + (672) + 2 ╳ (5556) + (25 600 + 12 288) + (25 600 + 12 288) + 6144= 126 024 bit times = 15.5KB

  * When it comes to 100G or above, or when cable length increases, such as data center interconnection, cable delay will be significant.

- Furthermore, implementation dependent internal buffer fragmentation should be considered when calculating headroom.
  - Buffer to store the packet is usually allocated chunk by chunk, not byte-stream FIFOs, e.g. 160 bytes as smallest chunk

Current PFC Headroom Reservation in Network Management is Not Efficient

- Usually, network engineers consider headroom during network deployment or network changes.
- One common way is to use default value from vendor. However, this rarely matches the real environment.
  - Variable distance impacting ‘cable delay’, especially in long distance Data Center Interconnect (DCI) scenario
  - Variable implementation dependent hardware processing impacting Interface Delay, Higher Layer Delay.



		Cable delay (bit times)
100G Base-R	10m	5000 (0.6KB)
	10km	5 000 000 (625KB)
0.6KB for 10m estimation error (DCN case); 625KB for 10km estimation error (DCI case)		

		ID + HD (bit times)
100G Base-R	802.3 max value	132 608
	Test value	100 000
Default settings may increase actual needs by 33%		

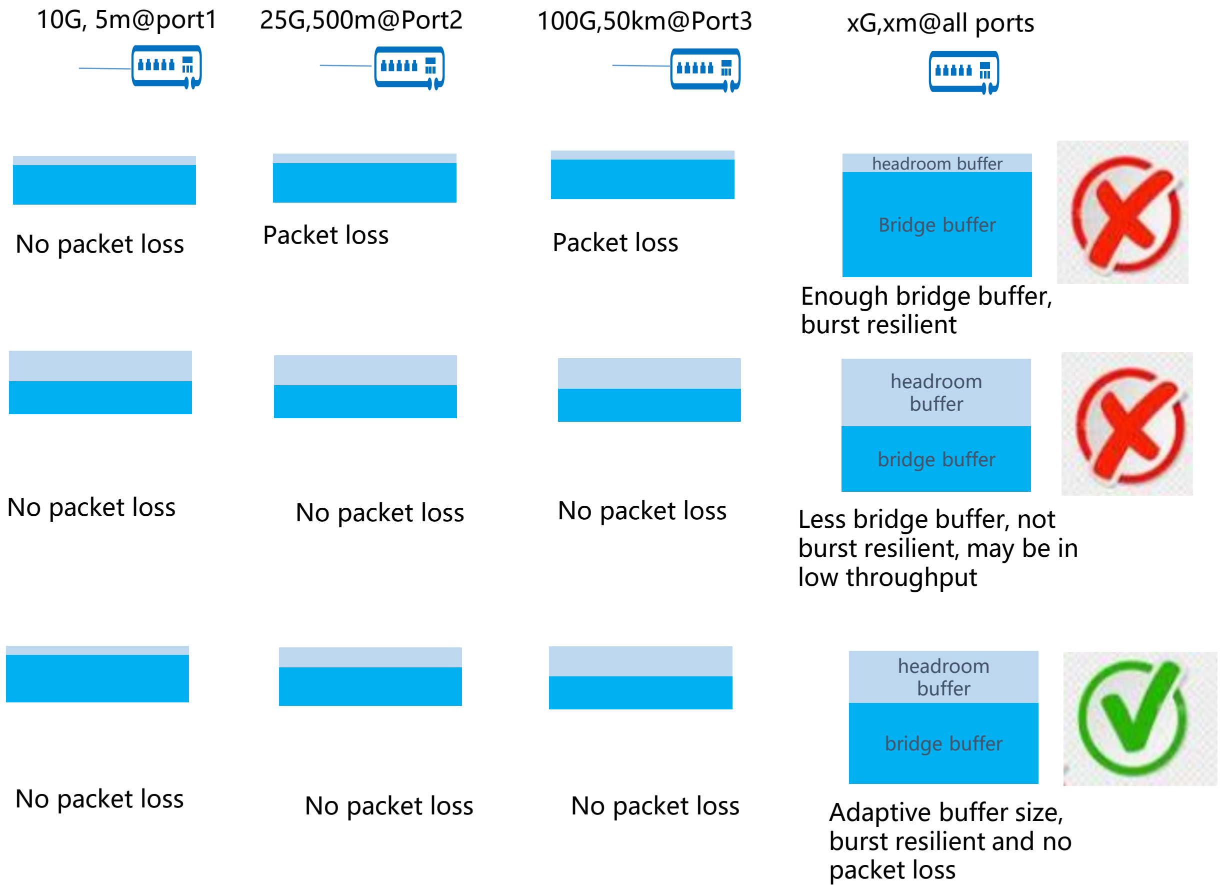
- Otherwise, manual calculation, configuration and test are required based on hop by hop distance, transmission rate, etc. That is time consuming.

Network Management Prefers 'Plug-and-Play'

As a network management engineer, I want to

- Place the switch wherever it is needed and assure lossless behavior.
- Not worry about an improper configuration (e.g default values) which might cause performance issue.
- Release me from learning complex operations (tools, commands, parameters, etc.) on different vendors' equipment, requiring me to read hundreds of pages of instructions.
- Shorten the network BIS(Bring into Service) time and reduce OPEX

If the headroom setting is automatically adapted to environment, like a 'plug and play' feature, the network manager's objectives can be met.



What We Have in IEEE 802

- ✓ 802.3 90 Ethernet support for time synchronization protocols--Data delay measurement

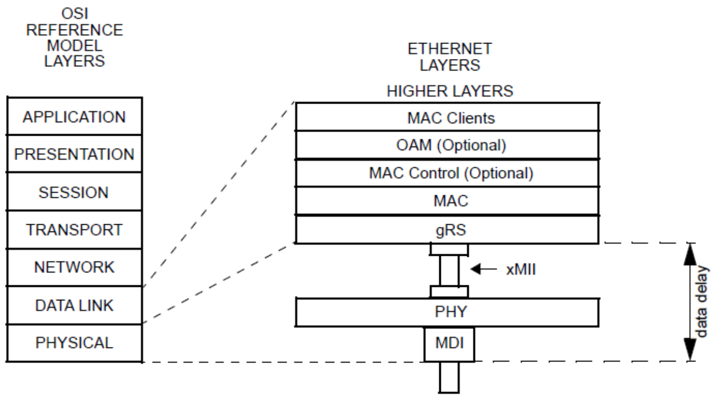
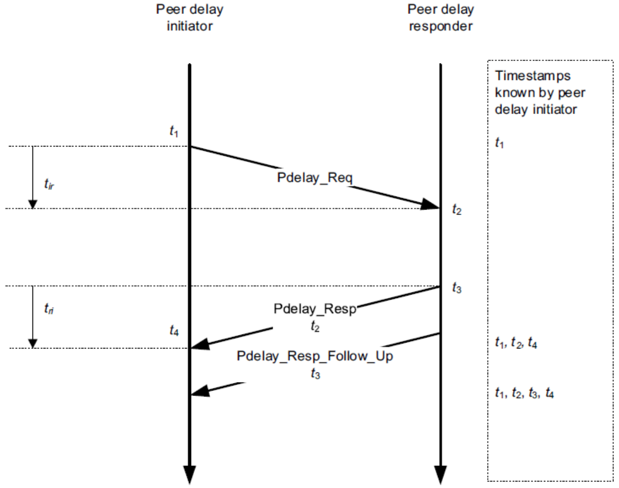


Figure 90–3—Data delay measurement

Measure tx/rx data path delay to support time synchronization
Data path delay is between xMII and MDI.

* Note: Current 802.3 timestamping is done at the xMII, 802.3cx propose to move the timestamp to MDI to improve time synch accuracy.

- ✓ 802.1AS 11.1.2 "Propagation delay measurement"



Use std 1588-2019 two-step PTP mechanism on a full-duplex point-to-point PTP link

- ✓ 802.1Qcc 12.32.2 "Propagation delay "

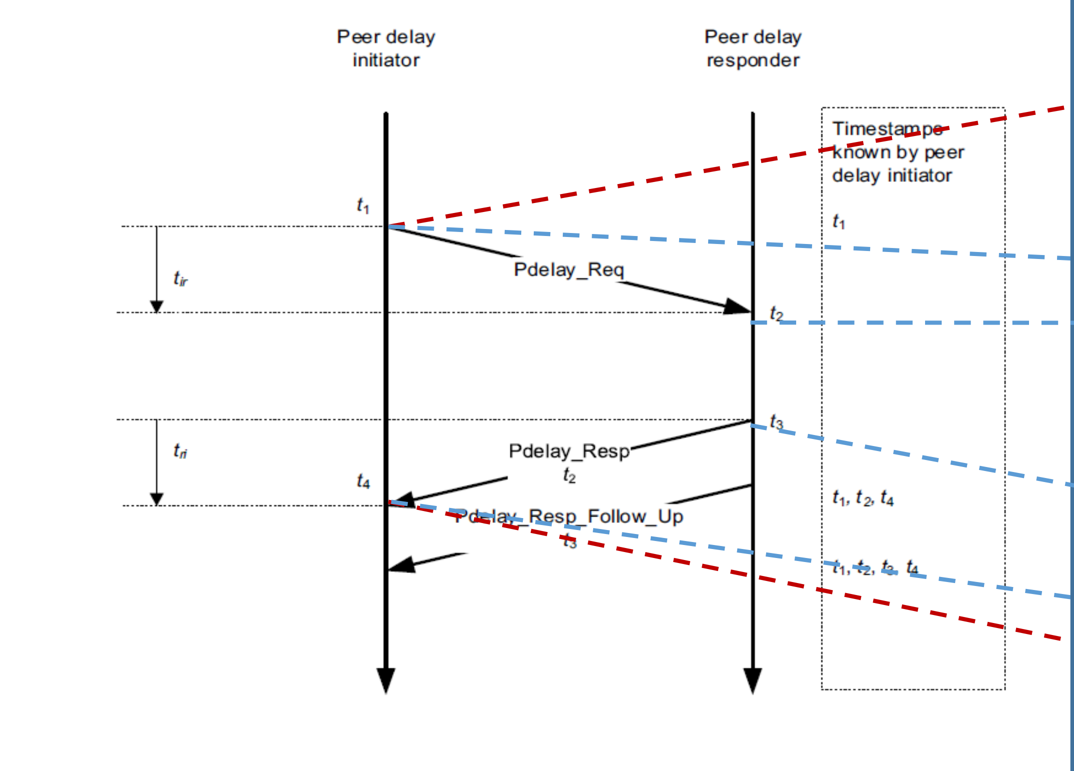
Name	Data type	Operations supported[a]	Conformance[b]	References
txPropagationDelay	unsigned integer	R	BE	12.32.2.1

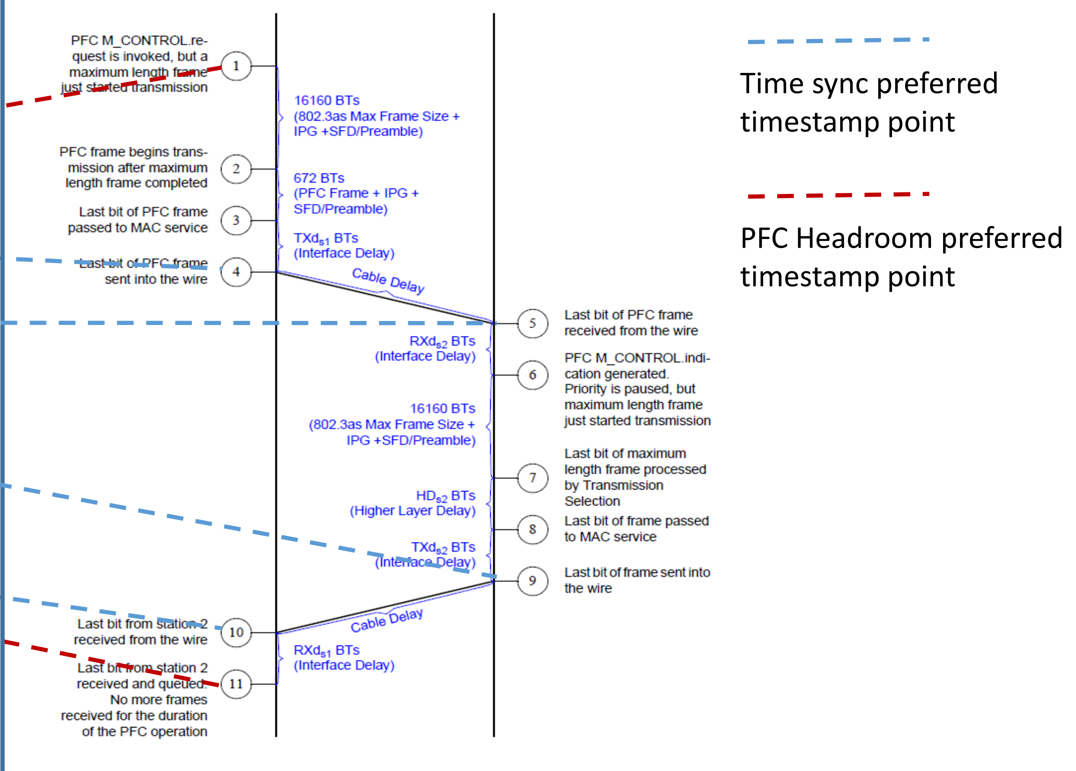
The txPropagationDelay attribute is typically measured using a time synchronization protocol, e.g. 802.1AS

What is Missing ……

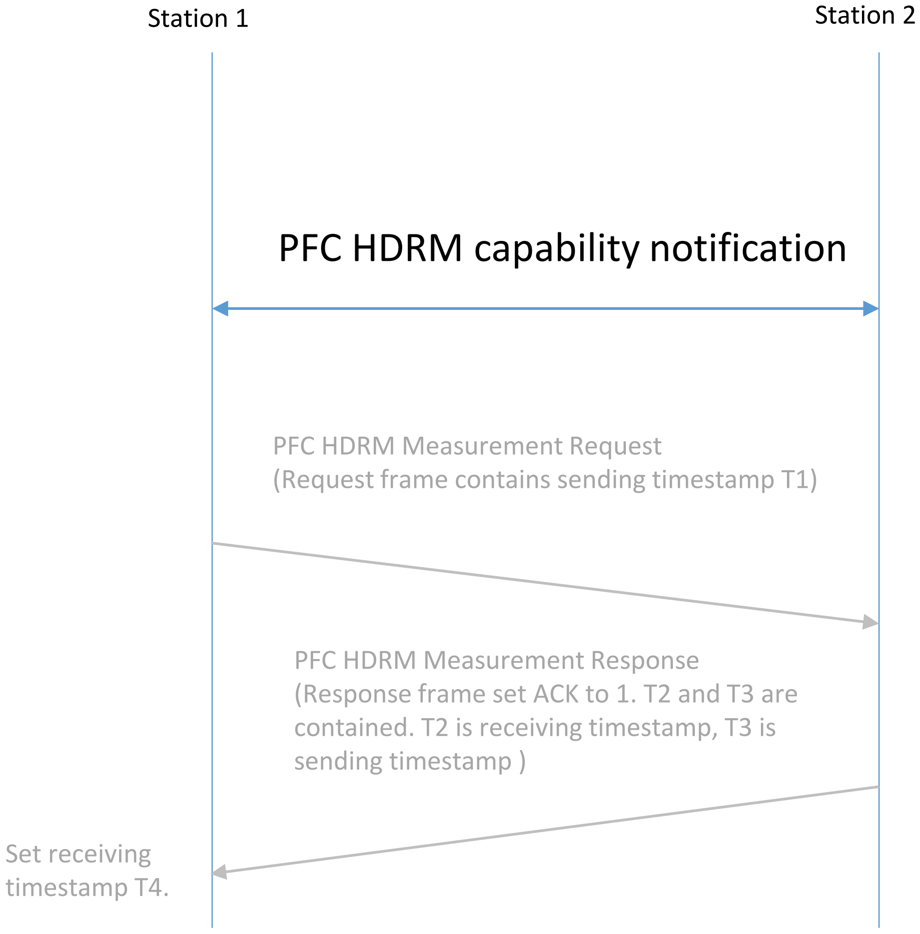
- The current set of standards are developed for time synchronization
  - Data center networks seldom activate time synchronization.
  - No description of usage for PFC headroom calculation.
  - Time synchronization has different preference of timestamp points from headroom calculation.

802.1 AS	PFC Headroom



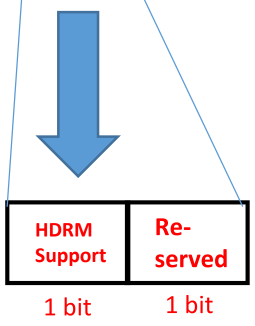


Proposal for Adaptive PFC Headroom (1/4)



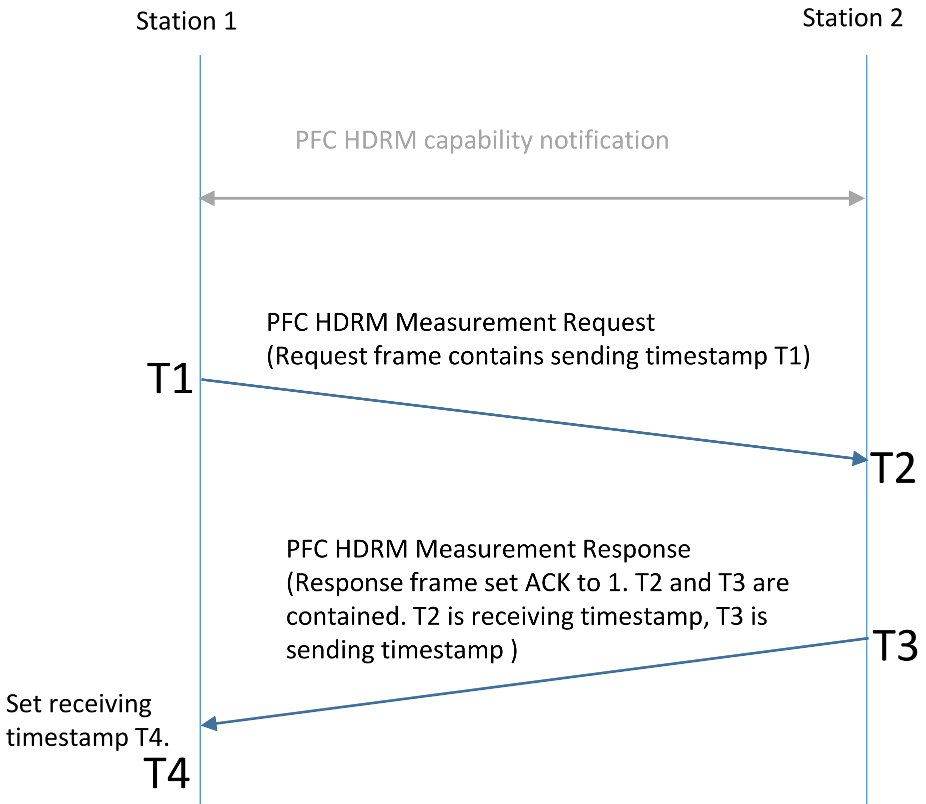
- Phase 1: Capability notification
  - Augment DCBX by extending PFC configuration TLV
  - DCBX uses LLDP with new PFC configuration TLV to exchange capability
  - If both support PFC HDRM, initiate PFC HDRM Measurement Request, otherwise, stop the procedure.

TLV type = 127	TLV information string length = 6	802.1 OUI 00-80-C2	802.1 subtype = 11	Willing	MBC	Re-served	PFC cap	PFC Enable
7 bits	9 bits	3 octets	1 octet	1 bit	1 bit	2 bits	4 bits	1 octet

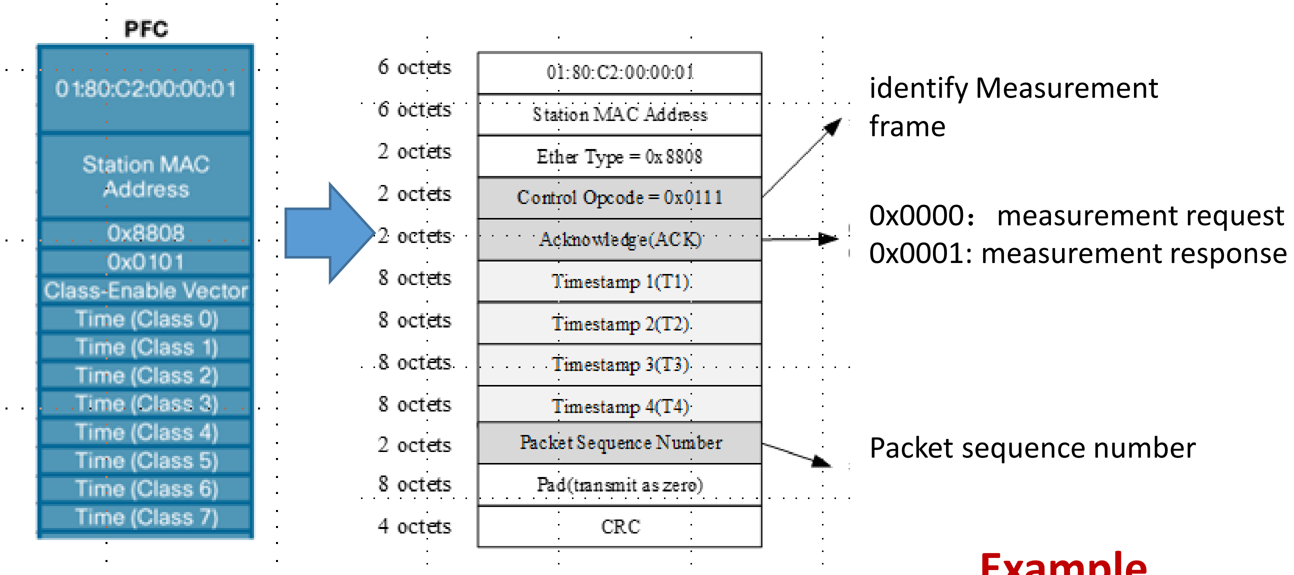


Example

Proposal for Adaptive PFC Headroom (2/4)



- Phase 2: Delay Measurement
  - Measurement request is sent from station 1 to station 2 with sending timestamp T1
  - Measurement response is sent from station 2 to station 2 with receiving timestamp T2 and sending timestamp T3
  - Station 1 set receiving timestamp T4
  - Measurement request and response frame is a new MAC control frame



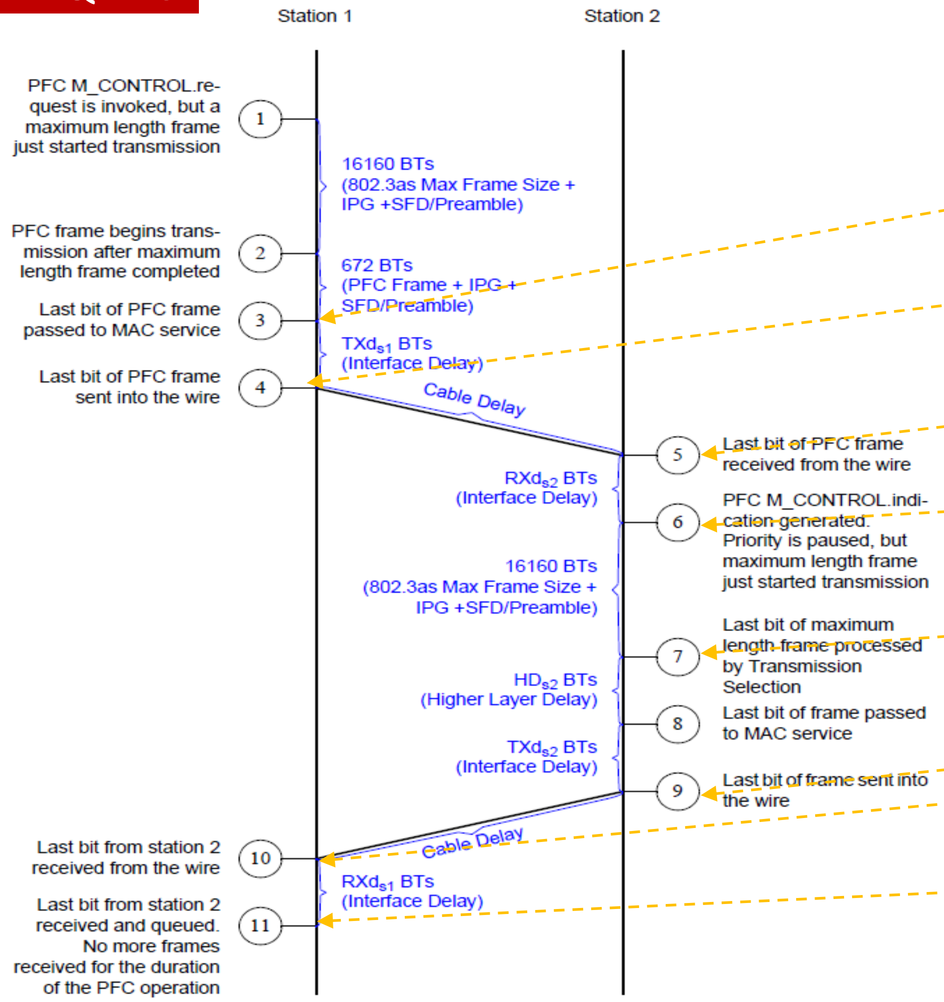
PFC frame format

Measurement frame format

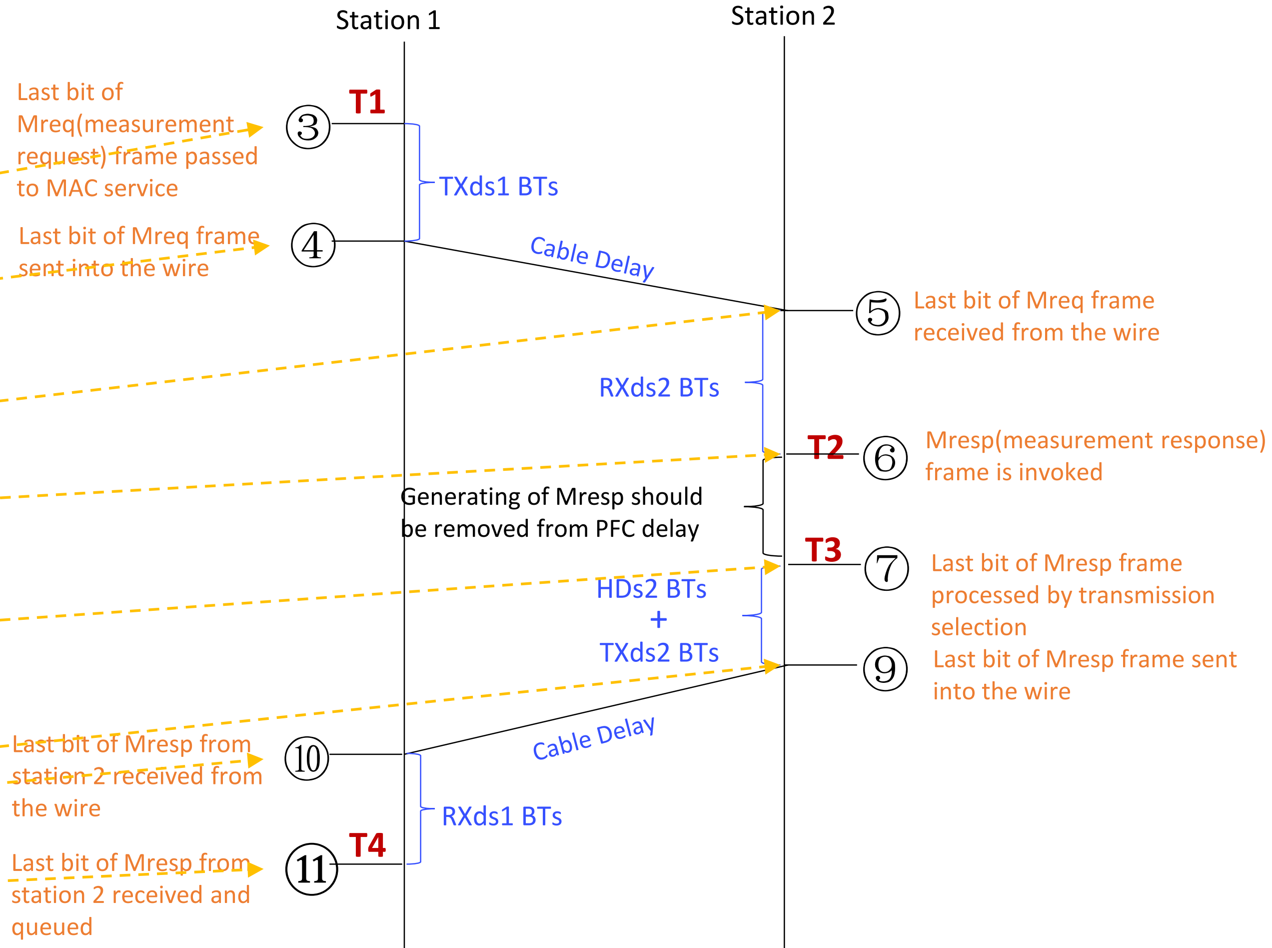
Example

Proposal for Adaptive PFC Headroom (3/4)

802.1Q PFC



Example of Adaptive headroom



DV = 2*(Max Frame) + (PFC Frame) + 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1

=

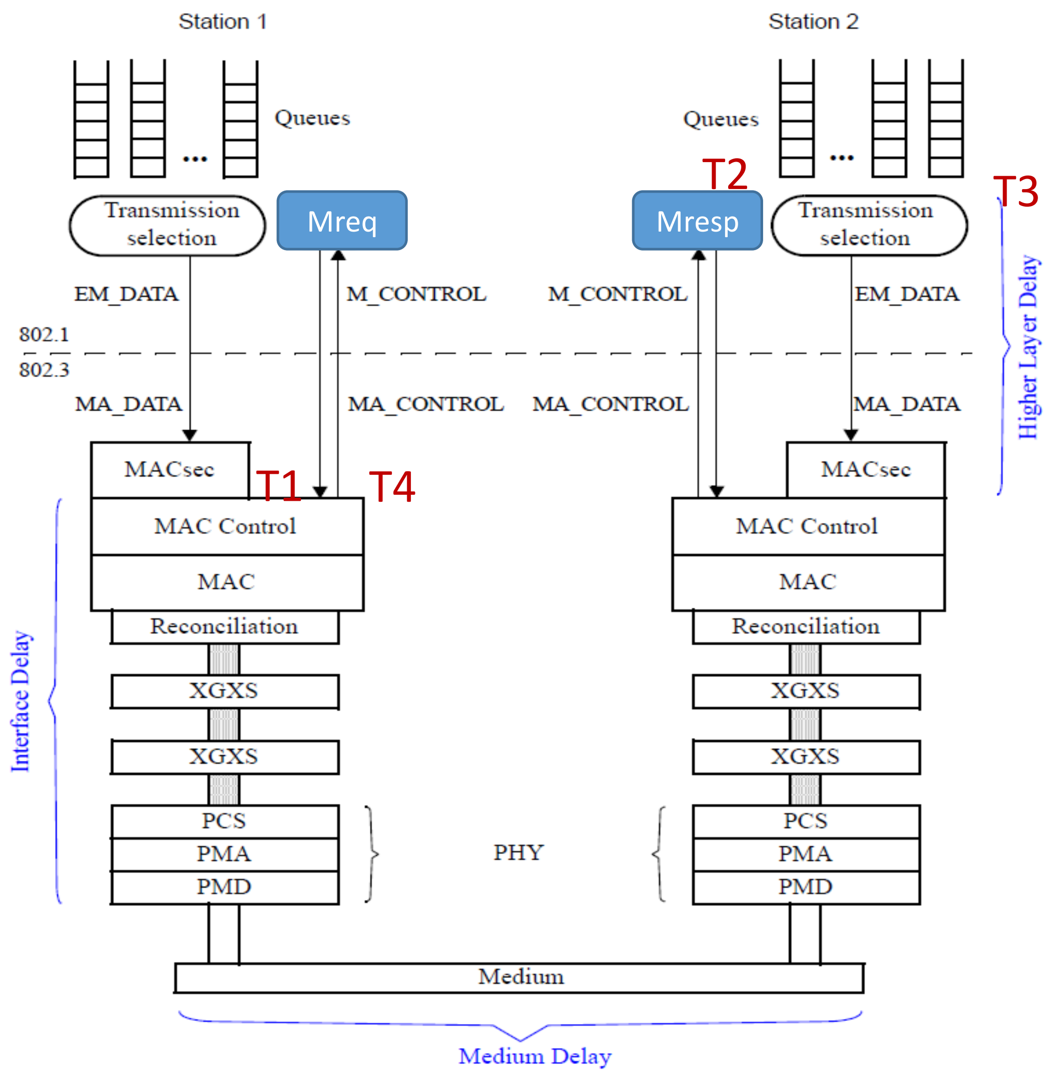
X = (T4-T1- (T3-T2)) * Speed = 2*(Cable Delay) + TXds1 + RXds2 + HDs2 + TXds2 + RXds1

DV = 2*(Max Frame) + (PFC Frame) + X

Proposal for Adaptive PFC Headroom (4/4)

Example

- Phase 3: Headroom calculation
  - X = Port speed * (T4-T1-(T3-T2))
  - DV = X + 2*(Max Frame) + (PFC Frame)
  - Headroom = DV * alpha
    - alpha is implementation dependent, considering buffer chunk size



Next Steps

- Define PFC headroom measurement mechanism
  - Measurement capability advertisement
  - Measurement timestamp point
  - Measurement frame interaction
  - Measurement frame format
  - Calculation method with timestamp
- Consideration of changes to 802.1Q

THANK YOU